

A Data-driven Situation Awareness Method Based on Random Matrix for Future Grids

Xing He, Lei Chu, Qian Ai, *Senior Member, IEEE*, Robert C. Qiu, *Fellow, IEEE*, Zenan Ling

Abstract—Data-driven methodologies are more suitable for a complex grid with readily accessible data when tasked with situation awareness. However, it is a challenge to turn the massive data, especially those with some spatial or temporal errors, into the driving force within tolerable cost of resources such as time and computation. This paper, based on random matrix theory (RMT), outlines a novel data-driven methodology. 1) Background information and previous work are reviewed. 2) Related to the methodology, the technical route and applied framework, data-proceeding and each procedure, evaluation system and related indicator set, and the advantages over classical methodologies are studied. Moreover, we make a comparison with the data-driven methodology based on Principal Component Analysis (PCA). 3) Related functions, including anomaly detection, spectrum test, correlation analysis, fault diagnosis and location, statistical indicator system and its visualization (i.e. 3D power map), are developed. This methodology gains insight into the large-scale interconnected grid in a more precise and natural way; it is model free requiring no knowledge about the physical model parameters. The methodology, in a flexible and holistic way, processes massive data in the form of large random matrix to depict a global but not a local picture of the system. Meanwhile, the large data dimension N and the large time span T , from the spatial aspect and the temporal aspect respectively, benefit the engineering performance of the proposed methodology; for this paper, the robustness against unsynchronized data is highlighted.

Index Terms—data analysis, random matrix theory, statistical indicator, data-driven, function, unsynchronized data

I. INTRODUCTION

SITUATION awareness (SA) is of great significant in power system operation, and a reconsideration of SA is essential for future grids [1]. Future grids are always huge in size and possess lots of flexible elements; besides, they operate under a novel regulation and their management mode is much different from prevailing one [2]. All these bring great challenges to or even disable the traditional model-based SA methods.

On the contrary, data are more and more easily accessible and the data-driven solution becomes an alternative to gain insight into the grid. The data-driven methodology faces some problems as follows:

- There are massive data in power grids. The dimensional curse [3] is inevitably generated and daily aggregated.
- Data are at low unit-value. The resource cost (time, hardware, human, etc.) for value extraction should be tolerable.
- Inevitably, for a massive data source there exist bad data, e.g. the incomplete, the inaccurate, the unsynchronized,

and the unavailable data. For system operations, decisions such as protection actions, should be highly reliable.

Data-driven approach and data utilization for smart grids are hotspots. We would like to introduce “Big Data Analytics for Grid Modernization” [4]—the Special Issue of IEEE Transaction on Smart Grid in Sep. 2016—as a clue. Several SA topics are discussed: References [5, 6] conduct anomaly detection and classification; Reference [7] takes into account social media such as Twitter; References [8, 9] study the estimation of the active ingredients such as PV Installations; Reference [10] utilizes the real-time data for online transient stability evaluation. Among these references, numerous algorithms/tools and frameworks/procedures are proposed as the key to the data-driven mode, such as Bayesian, principal component analysis (PCA) and singular value decomposition (SVD), support vector machine (SVM) and core vector machine (CVM), and matching pursuit decomposition (MPD). Besides this special issue, there are also numerous relevant work: References [11–13] show the improvement in wide-area monitoring, protection and control (WAMPAC) and the utilization of PMU data. References [14–16] study the fault detection and location. Xie et al., based on PCA, proposes an online application for early event detection by introducing a reduced dimensionality [17]; Lim et al., based on SVD, studies the quasi-steady-state operational problems relevant to the voltage instability phenomena [18].

A. Contribution

This paper, based on random matrix theory (RMT), proposed a data-driven methodology, which is aimed at effective data utilization. Section II gives the mathematical background and theoretical foundation. Spectrum test is introduced as a novel tool. Section III studies the details about the methodology, including the technical route and applied framework, data-proceeding and each procedure, evaluation system and related indicator set, and the advantages over classical methodologies. Specifically, a comparison with PCA-based methodology is made. Section IV, with concrete cases, presents the function designing and the unique advantages to validate the proposed methodology. The *spectrum test* is confirmed to be a competent anomaly detection tool *for the first time*. Particularly, the *robustness against unsynchronized data* is highlighted.

B. Previous Work

Paper [2] is our first relevant work; it is the first to propose the framework which introduces RMT into power systems. Ring Law and Marchenko-Pastur (M-P) Law are given as

the statistical foundation, and Mean Spectral Radius (MSR) is proposed as the high-dimensional indicator. Then we move forward to the second stage—paper [19] studies the correlation analysis under the above framework. Concatenated matrix \mathbf{A}_i is the key. It consists of the basic matrix \mathbf{B} and a factor matrix \mathbf{C}_i , i.e., $\mathbf{A}_i = [\mathbf{B}; \mathbf{C}_i]$. The LES indicators of these \mathbf{A}_i are computed in parallel to find out the sensitive factors. This study contributes to fault detection and location, line-loss reduction, and power-stealing prevention [20]. We also conduct analysis for power transmission equipment based on the same theoretical foundation [21]. Paper [1] is the third step in which LES set is discussed. Based on the LES set, a statistical indicator system, rather than a deterministic and model-based one, is built to help understand the system from a high-dimensional perspective. The robustness against data spatial errors and even data losses in the core area is emphasized.

II. MATHEMATICAL BACKGROUND AND THEORETICAL FOUNDATION

A. Random Matrix Modeling

Power grids operate in a balance situation obeying

$$\begin{cases} \Delta P_i = P_{is} - P_i(\mathbf{v}, \theta) \\ \Delta Q_i = Q_{is} - Q_i(\mathbf{v}, \theta) \end{cases}, \quad (1)$$

where P_{is} and Q_{is} are the power injections on node i , while $P_i(\mathbf{v}, \theta)$ and $Q_i(\mathbf{v}, \theta)$ are the injections of the network satisfying

$$\begin{cases} P_i = V_i \sum_{j=1}^n V_j (G_{ij} \cos \theta_{ij} + B_{ij} \sin \theta_{ij}) \\ Q_i = V_i \sum_{j=1}^n V_j (G_{ij} \sin \theta_{ij} - B_{ij} \cos \theta_{ij}) \end{cases}. \quad (2)$$

For simplicity, combining (1) and (2), we obtain

$$\mathbf{W}_0 = f(\mathbf{X}_0, \mathbf{Y}_0), \quad (3)$$

where \mathbf{W}_0 is the vector of power injections on nodes depending on P_{is}, Q_{is} . \mathbf{X}_0 is the system status variables depending on V_i, θ_i , while \mathbf{Y}_0 is the network topology parameters depending on B_{ij}, G_{ij} .

For a certain fluctuations, we formulate the system as

$$\mathbf{W}_0 + \Delta \mathbf{W} = f(\mathbf{X}_0 + \Delta \mathbf{X}, \mathbf{Y}_0 + \Delta \mathbf{Y}). \quad (4)$$

With Taylor Expansion, (4) is yielded as

$$\begin{aligned} \mathbf{W}_0 + \Delta \mathbf{W} &= f(\mathbf{X}_0, \mathbf{Y}_0) + f'_{\mathbf{X}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{X} + f'_{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{Y} \\ &\quad + \frac{1}{2} f''_{\mathbf{X}\mathbf{X}}(\mathbf{X}_0, \mathbf{Y}_0) (\Delta \mathbf{X})^2 + \frac{1}{2} f''_{\mathbf{Y}\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0) (\Delta \mathbf{Y})^2 \\ &\quad + f''_{\mathbf{X}\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{X} \Delta \mathbf{Y} + \dots \end{aligned} \quad (5)$$

Equ. (2) shows that \mathbf{W}_0 is linear with \mathbf{Y}_0 ; it means that $f''_{\mathbf{Y}\mathbf{Y}}(\mathbf{X}, \mathbf{Y}) = 0$. On the other hand, the value of system status variables \mathbf{X} are relatively stable and we can ignore $(\Delta \mathbf{X})^2$ and higher-order terms. In this way, we turn (5) into

$$\begin{aligned} \Delta \mathbf{W} &= f'_{\mathbf{X}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{X} + f'_{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{Y} \\ &\quad + f''_{\mathbf{X}\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{X} \Delta \mathbf{Y}. \end{aligned} \quad (6)$$

Suppose the network topology is unchanged, i.e., $\Delta \mathbf{Y} = 0$, from (6) we deduce that

$$\Delta \mathbf{X} = (f'_{\mathbf{X}}(\mathbf{X}_0, \mathbf{Y}_0))^{-1} \Delta \mathbf{W} = \mathbf{S}_0 \Delta \mathbf{W}. \quad (7)$$

On the other hand, suppose the power demands is unchanged, i.e., $\Delta \mathbf{W} = 0$, from (6) we deduce that

$$\Delta \mathbf{X} = \mathbf{S}_0 \Delta \mathbf{W}_y, \quad (8)$$

where $\mathbf{W}_y = [\mathbf{I} + f''_{\mathbf{X}\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0) \Delta \mathbf{Y} \mathbf{S}_0]^{-1} [f'_{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}_0)]$.

Note that $\mathbf{S}_0 = (f'_{\mathbf{X}}(\mathbf{X}_0, \mathbf{Y}_0))^{-1}$, i.e., the inversion of the Jacobian matrix \mathbf{J}_0 , expressed as

$$\mathbf{J}_{ij0} = \begin{bmatrix} \frac{\partial P_i}{\partial U_j} & \frac{\partial P_i}{\partial \theta_j} \\ \frac{\partial Q_i}{\partial U_j} & \frac{\partial Q_i}{\partial \theta_j} \end{bmatrix} \bigg|_{U_j=U_{j0}, \theta_j=\theta_{j0}}. \quad (9)$$

Thus, we describe the power system operation using random matrix—if there is an unexpected active power change or short circuit, the corresponding change of system status variables \mathbf{X}_0 , i.e. V_i, θ_i , will obey (7) or (8) respectively.

For a practical system, we can always build a relationship in the form of $\mathbf{Y} = \mathbf{H}\mathbf{X}$ with a similar procedure as (3) to (8); it is linear in high dimensions. Taking the case in [1] as an example, for an equilibrium operation system in which the reactive power is almost constant or changes much more slowly than the active one, the relationship model between voltage magnitude and active power, just like the Multiple Input Multiple Output (MIMO) model, is built as $\mathbf{V} = \mathbf{\Xi}\mathbf{P}$, as illustrated in Fig. 1. Note that most variables of vector \mathbf{V} are random due to the ubiquitous noises, e.g., small random fluctuations in \mathbf{P} . In addition, we can add very small artificial fluctuations to make them random or replace the missing/bad variables with random Gaussian variables. Furthermore, with the normalization, we can build the standard RMM in the form of $\tilde{\mathbf{V}} = \tilde{\mathbf{\Xi}}\mathbf{R}$, where \mathbf{R} is a standard Gaussian Random Matrix.

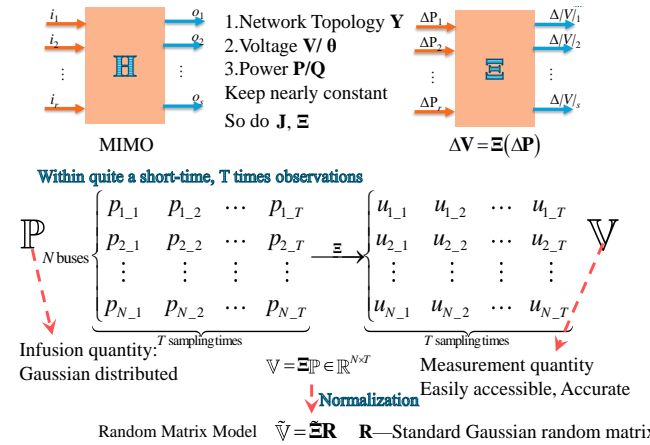


Fig. 1: Random Matrix Model for Power System Operation.

B. Anomaly Detection Based on Spectrum Analysis

In practice, Gaussian unitary ensemble (GUE) and Laguerre unitary ensemble (LUE) are mainly concerned as follows:

$$\mathbf{A} = \begin{cases} \frac{1}{2} (\mathbf{R} + \mathbf{R}^H) & , \mathbf{R} \in \mathbb{R}^{N \times N}, \text{GUE;} \\ \frac{1}{N} \mathbf{R} \mathbf{R}^H & , \mathbf{R} \in \mathbb{R}^{N \times T}, \text{LUE.} \end{cases}, \quad (10)$$

where \mathbf{R} is the standard Gaussian Random Matrix.

Let $p_{\mathbf{A}}(x)$ be the empirical density of \mathbf{A} , and define its empirical spectral distribution (ESD) $F_{\mathbf{A}}(x)$:

$$F_{\mathbf{A}}(x) = \frac{1}{N} \sum_{i=1}^N I_{\{\lambda_i \leq x\}}, \quad (11)$$

where \mathbf{A} is GUE or LUE matrix, $I(\cdot)$ represents the event indicator function. We investigate the rate of convergence of the expected ESD $\mathbb{E}\{F_{\mathbf{A}}(x)\}$ to the Wigner's Semicircle Law or Wishart's M-P Law.

Let $g_{\mathbf{A}}(x)$ and $G_{\mathbf{A}}(x)$ denote the empirical eigenvalue density and ESD of \mathbf{A} , and the Wigner's Semicircle Law and Wishart's M-P Law say:

$$g_{\mathbf{A}}(x) = \begin{cases} \frac{1}{2\pi} \sqrt{4 - x^2} & , x \in [-2, 2], \text{GUE}; \\ \frac{1}{2\pi cx} \sqrt{(x-a)(b-x)} & , x \in [a, b], \text{LUE}; \end{cases}, \quad (12)$$

where $a = (1 - \sqrt{c})^2, b = (1 + \sqrt{c})^2$.

$$G_{\mathbf{A}}(x) = \int_{-\infty}^x g_{\mathbf{A}}(u) du. \quad (13)$$

Then, we denote the Kolmogorov distance between $\mathbb{E}\{F_{\mathbf{A}}(x)\}$ and $G_{\mathbf{A}}(x)$ as Δ :

$$\Delta = \sup_x |\mathbb{E}\{F_{\mathbf{A}}(x)\} - G_{\mathbf{A}}(x)|. \quad (14)$$

Gotze and Tikhomirov, in their work [22], prove an optimal bound for Δ of order $O(N^{-1})$.

Lemma II.1. *There exists a positive constant C such that, for any $N \geq 1$,*

$$\Delta \leq CN^{-1}. \quad (15)$$

They also prove that the convergence of the density of standard Semicircle Law and M-P Law to the expected spectral density $p_{\mathbf{A}}(x)$ satisfies

Lemma II.2. *For GUE matrix, there exists a positive constant ε and C such that, for any $x \in [-2 + N^{-\frac{1}{3}}\varepsilon, 2 - N^{-\frac{1}{3}}\varepsilon]$,*

$$|p_{\mathbf{A}}(x) - g(x)| \leq \frac{C}{N(4 - x^2)}. \quad (16)$$

Lemma II.3. *For LUE matrix, let $\beta = N/T$, there exists some positive constant β_1 and β_2 such that $0 < \beta_1 \leq \beta \leq \beta_2 < 1$, for all $N \geq 1$. Then there exists a positive constant C and ε depending on β_1 and β_2 and for any $N \geq 1$ and $x \in [a + N^{-\frac{2}{3}}\varepsilon, b - N^{-\frac{2}{3}}\varepsilon]$,*

$$|p_{\mathbf{A}}(x) - h(x)| \leq \frac{C}{N(x-a)(b-x)}. \quad (17)$$

Lemma II.2 and II.3 also describe how fast the population distribution functions converge to the asymptotic limit. This spectrum test is interesting for anomaly detection about a large complicated grid; the effectiveness is validated in Section IV.

III. THE METHODOLOGY OF SITUATION AWARENESS

A. Technical Route and Practical Procedures

The proposed methodology consists of three essential procedures as illustrated in Fig. 2: 1) big data model—to model the system using experimental data for the RMM; 2) big data analysis—to conduct high-dimensional analyses for the indicator system as the statistical solutions; 3) engineering interpretation—to visualize and interpret the statistical results to human beings for decision-making.

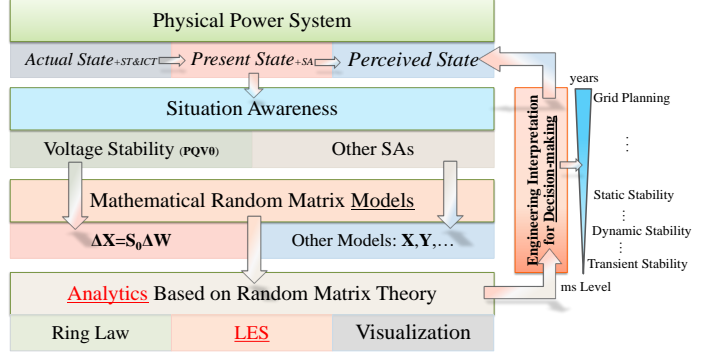


Fig. 2: SA Methodology based on RMT

This methodology is universal and adaptable to a wide variety of practical fields. We have already made numerous successful attempts in the field of anomaly detection and diagnosis for both the grid network [2, 19, 20] and the transmission equipment [21].

B. Classical Model-based Methodology

We would like to refer to Fig. 3 in book [23] as a clue. We are now entering the age of 4th-paradigm—data-intensive scientific discovery. Besides, the summaries for the classical decision-making approaches and for our proposed ones, obtained initially in [2], are improved in this paper as Fig. 4.

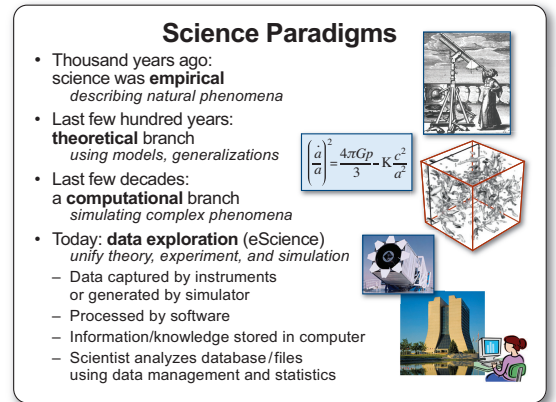


Fig. 3: Science Paradigms [24]

The second and third paradigms are typically model-based—they use equations, formulas, or simulations to describe the operation regulations and interaction mechanisms of a complicated system. The blue line in Fig. 4 depicts the general procedure, as well as corresponding tools. These tools

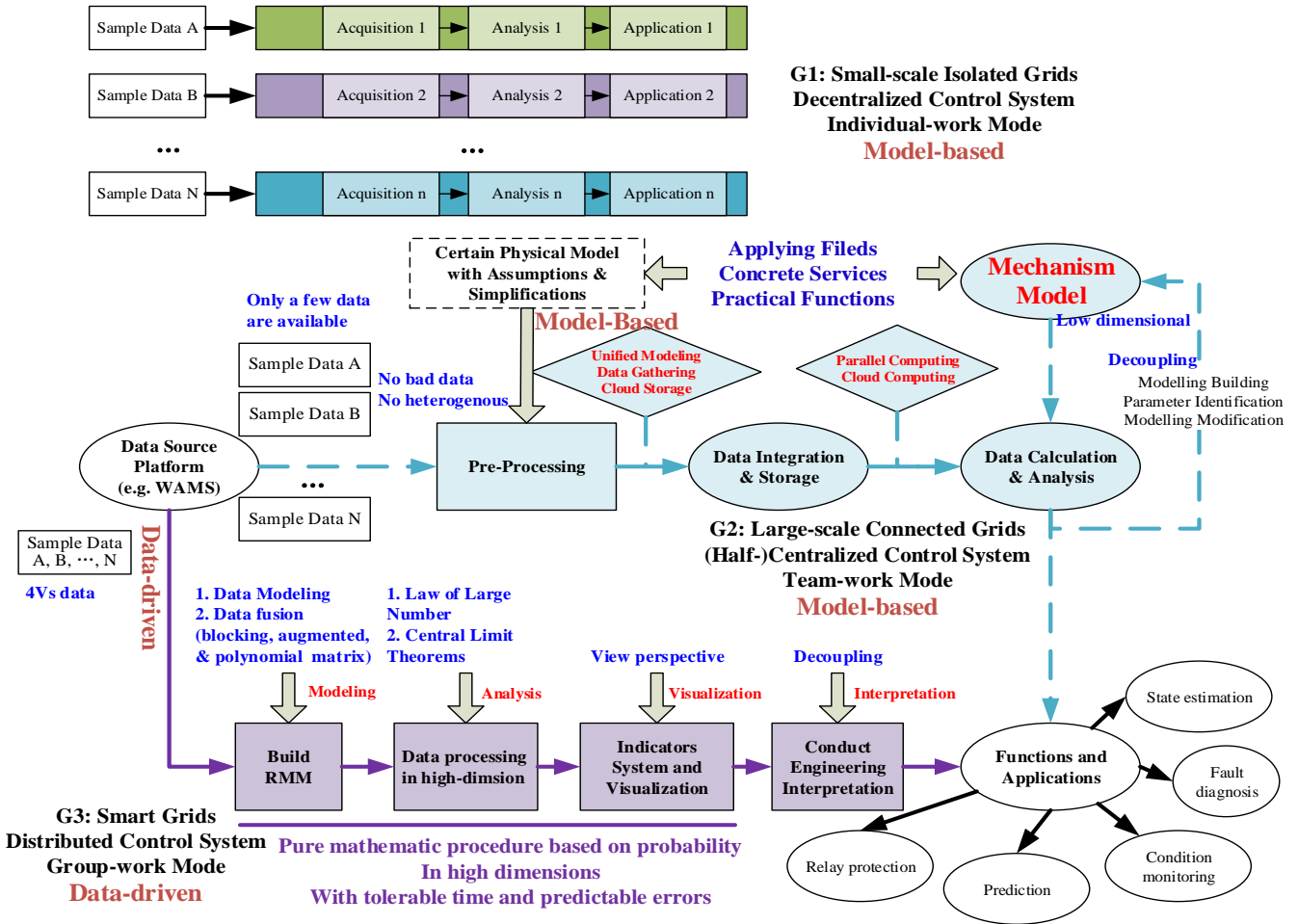


Fig. 4: Data utilization methodology for power systems. The above, middle, and below parts indicate the data processing procedures and the work modes for G1, G2, and G3, respectively.

cannot deal with massive data due to the essence of mechanism models—the models are in low dimensions and they only give a deterministic description. For instance, 1) when the initial value of PQ nodes, PV nodes, and the network topology are given, the grid's static state is obtained via the power flow equations, 2) ideally, wind power is proportional to the cube of its speed, and 3) the Lyapunov exponent which evaluates the transient stability is decided by the parameters of the grid network and the generating units. They all have one thing in common: deterministic indicators are employed as the key which are fully dependent upon only a few parameters. E.g., $y = ax^2 + bx + c$ is a 3-dimensional model—the relationship between x and y fully depends on a , b , and c .

Under classical statistical framework, only two typical data matrices in the form of $\mathbf{X} \in \mathbb{R}^{N \times T}$ are handleable: 1) N, T are small, and 2) N is small, T is infinite. It greatly restricts the utilization of the massive data; we should enable more data to speak for themselves [25]. In other words, model-based framework is not able to turn massive data into driven force. Although these massive data can contribute to model improvement and parameters correction, we can hardly conduct analysis more precisely with extremely large data volumes. Even worse, more data mean more errors; if we

take those bad data into the fixed model, poor results are obtained almost surely. Besides, the bias, caused by challenges such as error accumulations and spurious correlations, will not be alleviated via a low-dimensional procedure [19]; the dimensions of the procedure are limited by the dimensions of the model. The conclusion, data-driven mode is adapted to the future grid's analysis, is identified with the core viewpoint of the 4th-paradigm. The classical data utilization methodology is in urgent need of reform.

C. Classical Dimensionality Reduction Method—PCA

Data-driven Methodology is an alternative; it is model-free and able to process massive data in a holistic way. Principal component analysis (PCA) is one of the classical data processing algorithms which are sensitive to relative scaling original variables. It uses an orthogonal transformation to convert a set of possibly correlated raw variables into a set of linearly uncorrelated variables called principal components. The number of principal components is often much less than the number of original variables. In [17], PCA is used for dimensionality reduction from 14 PMU data sets to extract the event indicators. For PCA, the procedure consists of three parts: 1). Singular Value Decomposition (SVD) [18], 2).

Projection, and 3). Indicators.

1) *SVD*: The SVD of a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ is a factorization of the form $\mathbf{U}\mathbf{S}\mathbf{V}^H$, where $\mathbf{U}\mathbf{U}^H = \mathbf{E}_{T \times T}$ and $\mathbf{V}\mathbf{V}^H = \mathbf{E}_{N \times N}$. From \mathbf{U} , we choose the first m orthonormal vectors to form a matrix $\mathbf{U}_M \in \mathbb{R}^{T \times m}$ satisfying $\mathbf{U}_M^H \mathbf{U}_M = \mathbf{E}_{m \times m}$. Thus, the principal components space is constructed.

2) *Projection*: Consider a vector $\mathbf{b} \in \mathbb{R}^{T \times 1}$ and a subspace $\text{col}(\mathbf{A})$ ($\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m]$), we can obtain the projection matrix \mathbf{P} , projected vector \mathbf{p} , and the regression coefficients \mathbf{c} . In a similar way, for a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ and its principal components space $\text{col}(\mathbf{U}_M)$, we can push forward

$$\begin{aligned} \mathbf{P}\mathbf{b} = \mathbf{p} = \mathbf{A}\mathbf{c} & \Rightarrow \mathbf{P}\mathbf{X} = \mathbf{Z} = \mathbf{U}_M\mathbf{Y} \\ \left| \mathbf{P} = \mathbf{A}(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \right. & \Rightarrow \left| \mathbf{P} = \mathbf{U}_M \mathbf{U}_M^H \in \mathbb{R}^{T \times T} \right. \end{aligned} \quad (18)$$

3) *Indicators*: We can create various indicators with the training projection matrix \mathbf{P} , such as

$$\max \left([1 \ \cdots \ 1 \ 1]_{1 \times T} [\mathbf{P}\mathbf{X} - \mathbf{X}]_{T \times N} \right).$$

This procedure is applied to conduct early event detection; details can be found in [17]. Next, we will make a comparison between the PCA method and RMT method.

D. Data-driven Methodology—Random Matrix Theory

The procedure based on RMT is outlined below.

1) *Ring Law and MSR*: In contrast, Ring Law Analysis conducts SA as follows:

Steps of Ring Law Analysis

- 1) Select arbitrary raw data (or all available data) as data source Ω .
- 2) At a certain time t_i , form $\tilde{\mathbf{X}}$ as random matrix.
- 3) Obtain $\tilde{\mathbf{Z}}$ by matrix transformations ($\tilde{\mathbf{X}} \rightarrow \tilde{\mathbf{X}} \rightarrow \mathbf{X}_u \rightarrow \mathbf{Z} \rightarrow \tilde{\mathbf{Z}}$ [2]).
- 4) Calculate eigenvalues $\lambda_{\tilde{\mathbf{Z}}}$ and plot the Ring on the complex plane.
- 5) Conduct high-dimensional analysis.
 - 5a) Observe the experimental ring and compare it with the reference.
 - 5b) Calculate $\tau_{\text{MSR}} = \sum_{i=1}^N |\lambda_{\tilde{\mathbf{Z}},i}|/N$ as the *statistical indicators*.
 - 5c) Compare τ_{MSR} with the theoretical value $\mathbb{E}(\tau_{\text{MSR}})$.
- 6) Repeat 2)-5) at the next time point ($t_i = t_i + 1$).
- 7) Visualize τ on the time series, i.e. draw $\tau-t$ curve.
- 8) Make engineering explanations.

Steps 2–7, with a pure statistical procedure in high dimensions, conduct SA without any clue, assumption, or simplification. In step 2, arbitrary raw data, even ones from distributed nodes or intermittent time periods, are able to be focused on according to our will. The size of $\tilde{\mathbf{X}}$ is controllable, and the dimensionality curse is relieved in some ways.

2) *M-P Law and LES*: For the M-P Law Analysis, the steps are nearly the same, except as follows:

Partial Steps of M-P Law Analysis

- 3) Obtain \mathbf{M} by matrix transformations ($\mathbf{M} = \frac{1}{N} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^H$).
- 4) Calculate eigenvalues $\lambda_{\mathbf{M}}$.
- 5) Conduct high-dimensional analysis.
 - 5a) Observe the spectrum distribution and compare it with the reference.
 - 5b) Calculate $\tau = \sum_{i=1}^N \varphi(\lambda_{\mathbf{M},i})$ as the *statistical indicators*.
 - 5c) Compare τ with the theoretical value $\mathbb{E}(\tau)$.

E. Advantages of RMT-based Methodology

The data-driven methodology conducts analysis requiring no knowledge of system topologies, unit operation/control mechanism, causal relationship, etc. It is able to handle massive data all at once; large size of the data enhances the robustness of the final decision against the bad data (error, asynchronization, or loss). Comparing with classical data-driven methodologies (e.g. PCA), the RMT-based one has some unique characteristics:

1) The statistical indicator is generated from all the data as matrix entries but not chosen ones such as principal components. Thus, the RMT methodology is robust against classical data-driven methods' challenges such as error accumulations and spurious correlations [19].

2) For the statistical indicator, a theoretical or empirical value is able to be obtained in advance. The statistical indicator such as LES follows Gaussian distribution, and its bias is bounded [26] and often very small. For LES indicator, the bias is $O(n^{-2})$.

3) We can flexibly handle heterogenous data to realize data fusion via matrix operations, such as the blocking [2], the sum [27], the product [27], and the concatenation [19] of matrices. Data fusion is guided by the latest mathematical research [28, Chapter 7].

4) Only eigenvalues are used for further analyses, while the eigenvectors are omitted. This leads to much less required memory space and faster data-processing speed. Although some information is lost in this way, there is still lots of information contained in the eigenvalues [29], especially those outliers [30, 31].

5) Particularly, for a certain RMM $\hat{\mathbf{X}}$, diverse forms of LES are able to be constructed in parallel by designing different test functions without introducing any physical error (i.e. $\tau_F = \sum_{i=1}^N \varphi_F(\lambda_{\mathbf{M},i})$). Each LES, akin to a filter, provides a unique view-angle. Thus, the system is systematically understood piece by piece. Moreover, with a proper LES, we can trace some specific signal.

IV. CASE STUDIES

A. Background and Assumption of the Case

We adopt an IEEE 118-node system as the grid network (Fig. 14, same as [1]) and assume the events as Table I.

TABLE I: Series of Events

Stage	E1	E2	E3	E4
Time (s)	1–500	501–900	901–1300	1301–2500
$P_{\text{Node-52}}$ (MW)	0	$\uparrow 30$	$\uparrow 120$	$\nearrow t/4 - 205$

P_{52} is power demand of node 52.

The power demand of other nodes are assigned as

$$\tilde{y}_{\text{load_nt}} = y_{\text{load_nt}} \times (1 + \gamma_{\text{Mul}} \times r_1) + \gamma_{\text{Acc}} \times r_2, \quad (19)$$

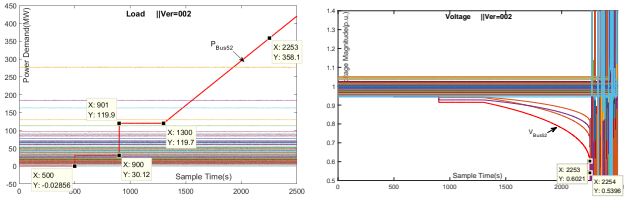
where r_1 and r_2 are the element of standard Gaussian Random Matrix; $\gamma_{\text{Acc}}=0.1$, $\gamma_{\text{Mul}}=0.001$. Thus, the power demand on each node is obtained as the system injections (Fig. 5a);

according to Fig. 1, the voltage is also obtained (Fig. 5b). Suppose we sample the voltage data at 1 Hz, the data source is formed denoted as $\Omega_V : \hat{v}_{i,j} \in \mathbb{R}^{118 \times 2500}$. The number of dimensions is $n = 118$ and the sampling time span is $t = 2500$.

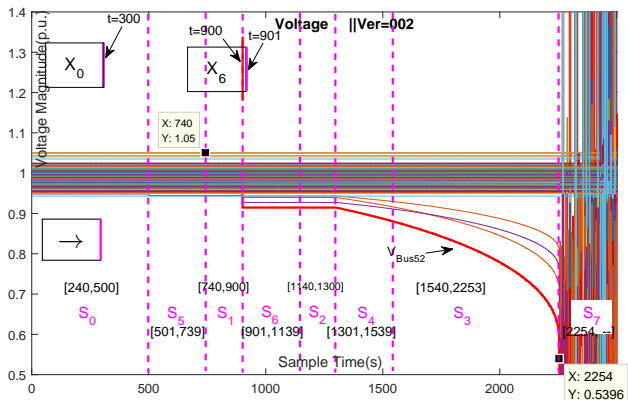
Suppose that the power demand data (Fig. 5a) is unknown or unqualified for SA due to the low sampling frequency or the bad quality. For further analysis, we just start with data source Ω_V (Fig. 5b) and assign the analysis matrix as $\mathbf{X} \in \mathbb{R}^{118 \times 240}$ (4 minutes' time span). First, we conduct category for the system operation status; the result are shown as Fig. 5c. In general, according to the raw data source and the analysis matrix size, we divide our system into 8 stages. Note that it is a statistical division—S4, S5, and S6 are transition stages, and their time span is right equal to the length of the analysis matrix minus ones, i.e., $T-1 = 239$. These stages are described as follows:

- For S0, S1, S2, the white noises play a dominant part. $P_{\text{Node-52}}$ is rising in turn.
- For S3, $P_{\text{Node-52}}$ maintains stable growth.
- S4, transition stage. Ramping signal exists.
- S5, S6, transition stages. Step signal exists.
- For S7, voltage collapse.

we also pick out two typical data cross-sections for stage S0 and S6: $\mathbf{X}_0 \in \mathbb{R}^{118 \times 240}$ during period $t = [61 : 300]$ at the sampling time $t_{\text{end}} = 300$, and 2) $\mathbf{X}_6 \in \mathbb{R}^{118 \times 240}$ during period $t = [662 : 901]$ at the sampling time $t_{\text{end}} = 901$.



(a) Assumed Event, Unavailable. (b) Raw Voltage, Ω_V for Analysis.



(c) Category for Operation Status and Selected Matrix Based on Ω_V .

Fig. 5: Assumed Event, Data Source, and Category for Case.

B. Anomaly Detection

1) Based on Ring Law and M-P Law:

According to our previous work [2], we build up the random matrix $\hat{\mathbf{V}}$ from the raw voltage data. Then, based on RMT,

we employ τ_{MSR} as a statistical indicator to conduct anomaly detection. For the selected data cross-section \mathbf{X}_0 and \mathbf{X}_6 , their M-P Law and Ring Law Analysis are shown as Fig 6a, 6b, 6c and 6d. With moving slide-window (MSW) technology, the $\tau_{\text{MSR}}-t$ curve is obtained as Fig. 6e.

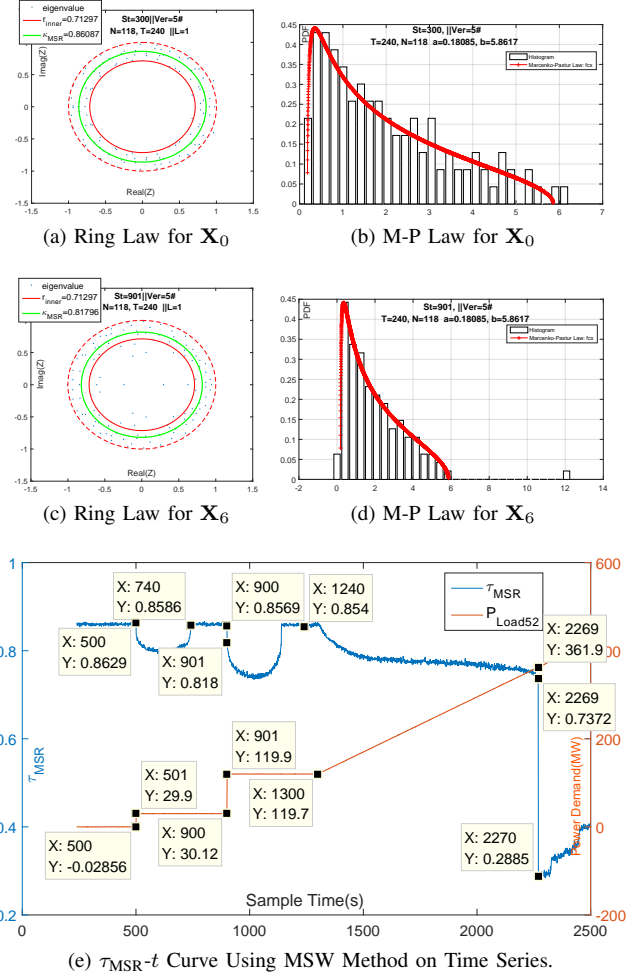


Fig. 6: Initial Anomaly Detection Result.

Fig 6 shows that when there is no signal in the system, the experimental RMM well matches the Ring Law and M-P Law, and the experimental value of LES is approximately equal to the theoretical value. Vice versa, at the very beginning ($t_{\text{end}} = 901$) of the step signal, the Ring Law and M-P Law are violated. Besides, the proposed high-dimensional indicator τ_{MSR} , is much more sensitive to the anomaly. τ_{MSR} starts the dramatic change as shown in the $\tau_{\text{MSR}}-t$ curve as Fig. 6e, while the raw voltage magnitudes are still in the normal range as shown in Fig. 5c. Moreover, following the previous work [1] we design numerous kinds of LES τ and make $\mu_0 = \tau / \mathbb{E}(\tau)$. The results are shown in Fig. 7 and prove that different indicators have different characteristics and effectiveness; it is another topic and we will not further discuss here.

2) Based on Spectrum Test:

We still set the sampling time at $t_{\text{end}} = 300$ and $t_{\text{end}} = 901$. Following Lemma II.2 and Lemma II.3, $\mathbf{Y}_0, \mathbf{Y}_6 \in \mathbb{R}^{118 \times 240}$ (span $t = [61 : 300]$ and $t = [662 : 901]$), and $\mathbf{Z}_0, \mathbf{Z}_6 \in \mathbb{R}^{118 \times 118}$ (span $t = [183 : 300]$ and $t = [784 : 901]$) are selected.

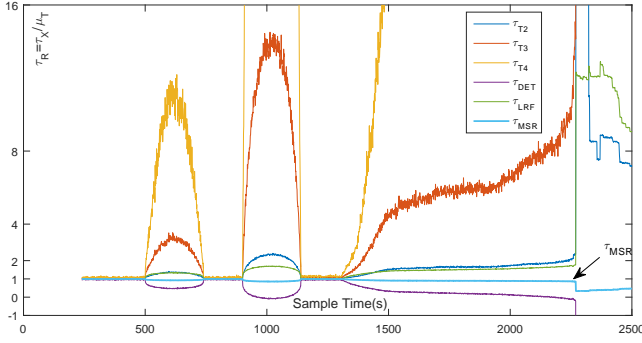


Fig. 7: Illustration of Various LES Indicators.

The results are shown in Fig. 8 and Fig. 9. These results validate that spectrum test is competent to conduct anomaly detection—when the power grid is under normal condition, the empirical eigenvalue density $p_A(x)$ and the ESD function $F_A(x)$ are almost strictly bounded between the upper bound and lower bound, and vice versa. Moreover, these results also validate that GUE and LUE are proper mathematical tools to model the power grid operation.

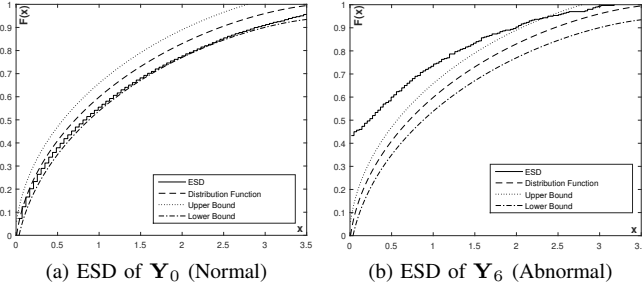


Fig. 8: Anomaly Detection Using LUE matrices

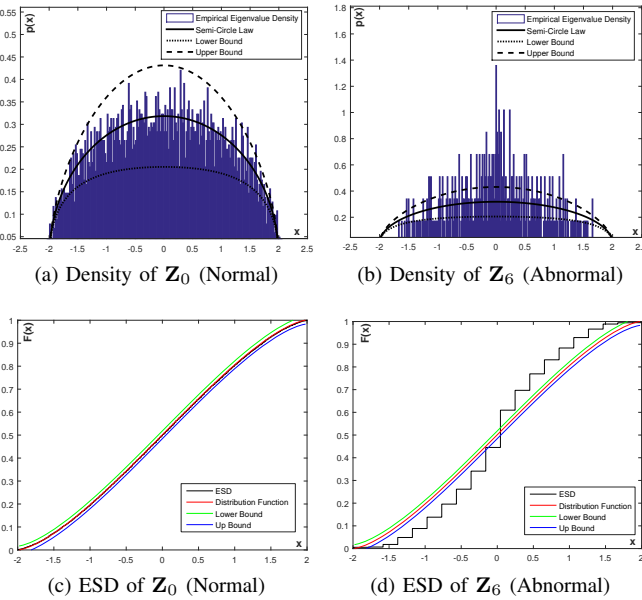


Fig. 9: Anomaly Detection Using GUE matrices

C. Steady Stability Evaluation

The $V - P$ curve (also called nose curve) and the smallest eigenvalue of the Jacobian Matrix [18] are the two clues for steady stability evaluation. In this case, we focus on the E4 part during which $P_{\text{Node-52}}$ keeps increasing to break down the steady stability. The related $V - P$ curve and $\lambda - P$ curve, respectively, are given in Fig. 10a and Fig. 10b. Only using the data source Ω_V , we choose some data cross-section, $T_1 : [1601 : 1840]$; $T_2 : [1901 : 2140]$; $T_3 : [2101 : 2340]$, as shown in Fig. 10a. The RMT-based results are shown as Fig. 11. The outliers become more evidence as the stability degree decreases. The statistic of the outliers is similar to the smallest eigenvalue of Jacobian Matrix, Lyapunov Exponent or the entropy in some sense.

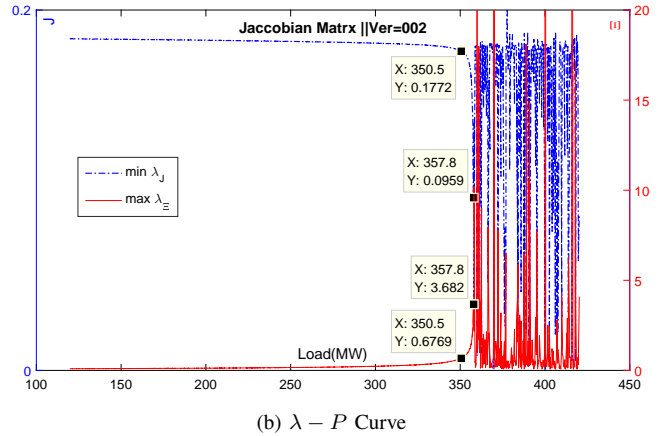
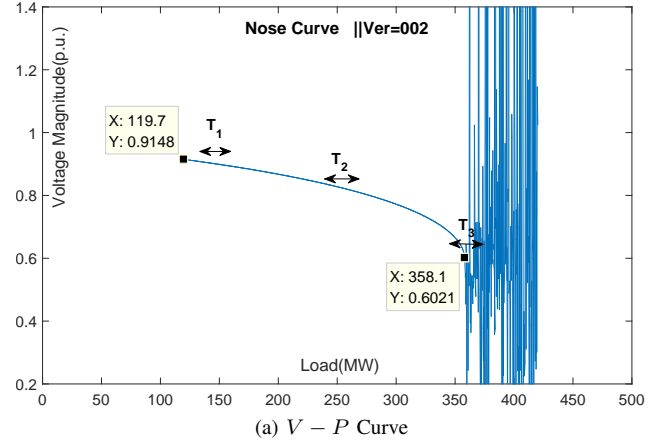


Fig. 10: The $V - P$ curve and $\lambda - P$ curve.

For further analysis, we take the signal and stage division into account. Generally speaking, sorted by the stability degree, the stages are ordered as $S0 > S1 > S2 > S3 \gg \max(S4, S5) \gg S6 \gg S7$. According to Fig. 7, we make the Table II. The high-dimensional indicators $\bar{\tau}_{\mathbf{X}_R}$ and V_R have the same trend with the stability degree order. These statistics have the potential for data-driven stability evaluation.

D. Correlation Analysis

The key for correlation analysis is the concatenated matrix A_i , which consist of two part—the basic matrix B and a

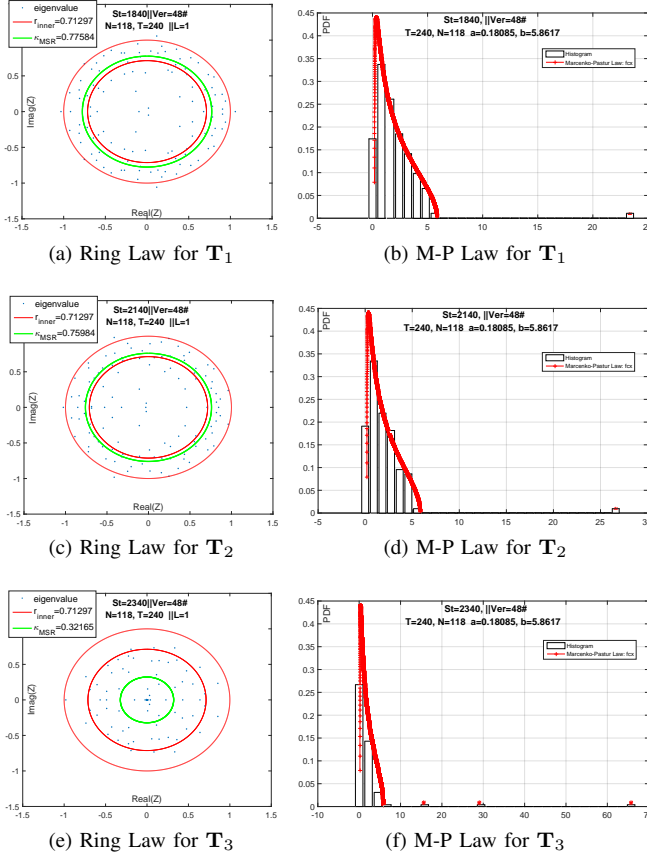


Fig. 11: RMT-based Results for Voltage Stability Evaluation.

TABLE II: Indicator of Various LESs at Each Stage.

	MSR	T ₂	T ₃	T ₄	DET	LRF
E₀: Theoretical Value						
$\mathbb{E}(\tau)$	0.8645	1338.3	10069	8.35E4	48.322	73.678
$\mathbb{D}_T(\tau)$	—	665.26	93468	1.30E7	1.3532	1.4210
S0 [0240:0500, 261]: Small fluctuations around 0 MW						
$\overline{\tau}_{\mathbf{X}_R}$	0.995	1.010	1.040	1.080	0.959	1.014
V	6E-6	78.38	3.03E4	7.14E6	0.4169	0.3908
V_R	1	1	1	1	1	1
S5 [0501:0739, 239]: A step signal (0 MW \uparrow 30 MW) is included						
$\overline{\tau}_{\mathbf{X}_R}$	0.9331	1.280	2.565	7.661	0.5453	1.284
V_R	1.49E1	1.64E2	1.16E3	8.63E3	3.43E1	3.97E1
S1 [0740:0900, 161]: Small fluctuations around 30 MW						
$\overline{\tau}_{\mathbf{X}_R}$	0.9943	1.010	1.039	1.084	0.9568	1.015
V_R	0.8608	0.9121	0.9476	1.234	0.8972	1.101
S6 [0901:1139, 239]: A step signal (30 MW \uparrow 120 MW) is included						
$\overline{\tau}_{\mathbf{X}_R}$	0.8742	2.054	1.06E1	7.22E1	7E-2	1.597
V_R	5.49E1	2.06E3	3.87E4	8.54E5	1.52E2	1.62E2
S2 [1140:1300, 161]: Small fluctuations around 120 MW						
$\overline{\tau}_{\mathbf{X}_R}$	0.9930	1.019	1.067	1.135	0.9488	1.021
V_R	0.7823	1.053	1.189	1.135	0.7310	0.9255
S4 [1301:1539, 239]: A ramp signal (119.7 MW \nearrow) is included						
$\overline{\tau}_{\mathbf{X}_R}$	0.9337	1.295	2.787	9.615	0.5316	1.294
V_R	8.50E1	7.41E2	5.63E3	5.17E4	2.14E2	2.30E2
S3 [1540:2253, 714]: Steady increase (\nearrow 358.1 MW)						
$\overline{\tau}_{\mathbf{X}_R}$	0.8906	1.717	6.530	3.48E1	0.1483	1.545
V_R	1.35E1	3.28E2	5.33E3	1.10E5	6.11E1	6.85E1
S7 [2254:2500, 247]: Static voltage collapse (361.9 MW \nearrow)						
$\overline{\tau}_{\mathbf{X}_R}$	0.4259	1.02E1	2.11E2	4.65E3	-1.4E1	1.08E1
V_R	1.94E3	5.81E5	1.20E8	3.2E10	9.02E4	9.62E4

$$*\overline{\tau}_{\mathbf{X}_R} = \overline{\tau}_{\mathbf{X}}/\mathbb{E}(\tau); V_R(\tau_{\mathbf{X}}) = V(\tau_{\mathbf{X}})/V(\tau_{\mathbf{X}_0}).$$

certain factor matrix C_i , i.e., $A_i = [B; C_i]$. For more details, see our previous work [19]. The LES of each A_i is computed in parallel, and Fig. 12 shows the results.

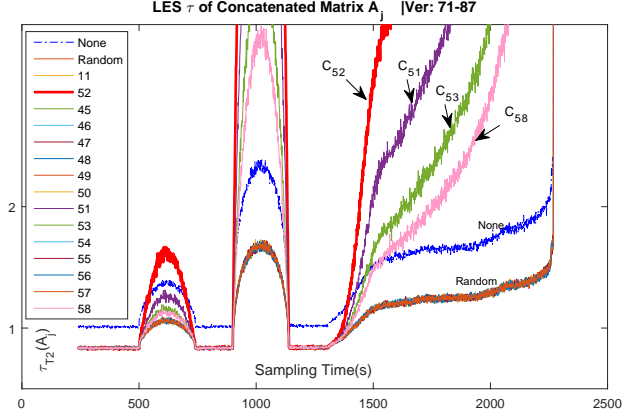


Fig. 12: Sensitivity Analysis based on Concatenated Matrix.

In Fig. 12, the blue dot line (marked with None) shows the LES of basic matrix B , and the orange line (marked with Random) shows the LES of the concatenated matrix $[B; R]$ (R is the standard Gaussian Random Matrix). Fig. 12 demonstrates that: 1) node 52 is the causing factor of the anomaly; 2) sensitive nodes are 51, 53, and 58; and 3) nodes 11, 45, 46, etc., are not affected by the anomaly. Based on this algorithm, we can continue to conduct behavior analysis, e.g., detection and estimation of residential PV installations [8]. Behavior analysis is a big topic. Limited to the space, we will not expand it here.

E. SA with Unsynchronized Data

The proposed data-driven method is robust against bad data in both spatial aspect and temporal one. In our previous work [1], we successfully conducted SA with data loss in the core area. This part we talk about SA with unsynchronized data. The phenomenon that unsynchronized data exists in the data platform such as SCADA and WAMS is very common. It is mainly caused by erroneous time-tags or communication delay. Sometimes, for a certain signal, the proper delay protection or causality/linkage/interaction/response mechanism will also lead to unsynchronized data. It is hard to measure or even detect the time delay via routine methods.

With the simulation data, we make an artificial 25 sampling points delay for 7 nodes—11, 14, 50, 52, 53, 77, and 81. With the concatenation operation introduced above, similarly, we obtain the results shown as Fig. 13. It is an interesting discovery that the methodology is robust against unsynchronized data: 1) the anomalies are detected at $t = 501$ and $t = 901$; 2) node 52 is the most sensitive node; 3) with more detailed observation, we can even quantitatively deduce that there exists a 25 sampling points delay (925 – 900) for node 52.

V. CONCLUSION

This paper summarizes and improves the previous work about power system big data analytics. It systematically proposed a data-utilization framework from the perspective of

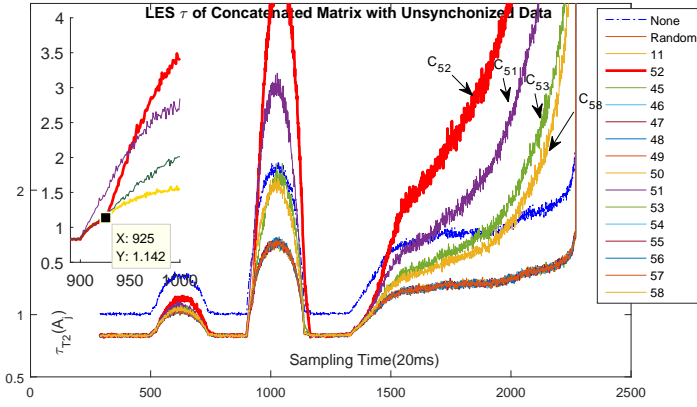


Fig. 13: Situation Awareness with Unsynchronized Data.

mathematics, engineering, and methodology. Random matrix theory (RMT) is presented as the theoretic foundation, and spectrum test is introduced as a novel SA tool.

The proposed RMT-based methodology has numerous unique advantages, and it is more suitable for complicated systems with easily accessible data. In the form of large random matrix, it handles massive data which are in high-dimension and within a wide time span all at once. In this way, highly reliable decisions are still attainable with some bad data, e.g., the unsynchronized data caused by erroneous time-tags or communication delays. Moreover, with the statistical processing such as test function setting, the proposed data-driven methodology has the potential to balance the perspectives of the speed, the sensitivity, and the reliability in practice.

The stability evaluation and behavior analysis are two big topics along this direction. Besides, the statistical indicators are good medium for artificial intelligence and machine learning.

APPENDIX A

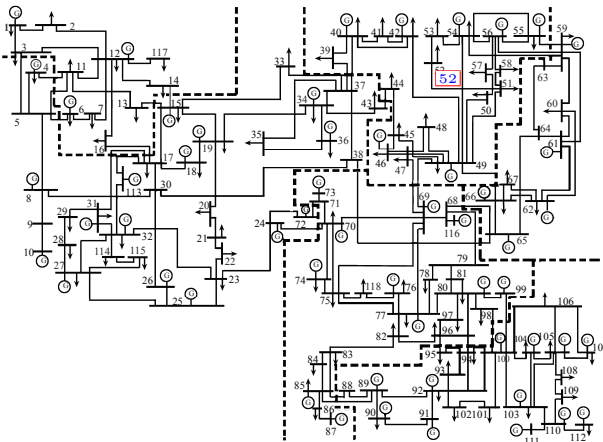


Fig. 14: Partitioning network for the IEEE 118-node system.

REFERENCES

- [1] X. He, R. C. Qiu, Q. Ai, L. Chu, and X. Xu, "Linear eigenvalue statistics: An indicator ensemble design for situation awareness of power systems," *ArXiv e-prints*, Dec. 2015. [Online]. Available: <http://arxiv.org/pdf/1512.07082.pdf>
- [2] X. He, Q. Ai, R. C. Qiu, W. Huang, L. Piao, and H. Liu, "A big data architecture design for smart grids based on random matrix theory," *ArXiv e-prints*, Jan. 2015, accepted by *IEEE Trans on Smart Grid*. [Online]. Available: <http://arxiv.org/pdf/1501.07329.pdf>
- [3] L. Moulin, A. Alves da Silva, M. El-Sharkawi, R. J. Marks *et al.*, "Support vector machines for transient stability analysis of large-scale power systems," *Power Systems, IEEE Transactions on*, vol. 19, no. 2, pp. 818–825, 2004.
- [4] T. Hong, C. Chen, J. Huang, N. Lu, L. Xie, and H. Zareipour, "Guest editorial big data analytics for grid modernization," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2395–2396, Sept 2016.
- [5] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window pca," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2537–2548, Sept 2016.
- [6] H. Jiang, X. Dai, D. W. Gao, J. J. Zhang, Y. Zhang, and E. Muljadi, "Spatial-temporal synchrophasor data characterization and analytics in smart grid fault detection, identification, and impact causal analysis," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2525–2536, Sept 2016.
- [7] H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2516–2524, Sept 2016.
- [8] X. Zhang and S. Grijalva, "A data-driven approach for detection and estimation of residential pv installations," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2477–2485, Sept 2016.
- [9] H. Shaker, H. Zareipour, and D. Wood, "A data-driven approach for estimating the power generation of invisible solar sites," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2466–2476, Sept 2016.
- [10] B. Wang, B. Fang, Y. Wang, H. Liu, and Y. Liu, "Power system transient stability assessment based on big data and the core vector machine," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2561–2570, Sept 2016.
- [11] A. Phadke and R. M. de Moraes, "The wide world of wide-area measurement," *Power and Energy Magazine, IEEE*, vol. 6, no. 5, pp. 52–65, 2008.
- [12] V. Terzija, G. Valverde, D. Cai, P. Regulski, V. Madani, J. Fitch, S. Skok, M. M. Begovic, and A. Phadke, "Wide-area monitoring, protection, and control of future electric power networks," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 80–93, 2011.
- [13] L. Xie, Y. Chen, and H. Liao, "Distributed online monitoring of quasi-static voltage collapse in multi-area power systems," *Power Systems, IEEE Transactions on*, vol. 27, no. 4, pp. 2271–2279, 2012.
- [14] Q. Jiang, X. Li, B. Wang, and H. Wang, "Pmu-based fault location using voltage measurements in large transmission networks," *Power Delivery, IEEE Transactions on*, vol. 27, no. 3, pp. 1644–1652, 2012.
- [15] M. Venugopal and C. Tiwari, "A novel algorithm to determine fault location in a transmission line using pmu measurements," in *Smart Instrumentation, Measurement and Applications (IC-SIMA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–4.
- [16] A. H. Al-Mohammed and M. Abido, "A fully adaptive pmu-based fault location algorithm for series-compensated lines," *Power Systems, IEEE Transactions on*, vol. 29, no. 5, pp. 2129–2137, 2014.
- [17] L. Xie, Y. Chen, and P. Kumar, "Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis," *Power Systems, IEEE Transactions on*, vol. 29, no. 6, pp. 2784–2794, 2014.
- [18] J. M. Lim and C. L. DeMarco, "Svd-based voltage stability assessment from phasor measurement unit data," *IEEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1–9, 2015.
- [19] X. Xu, X. He, Q. Ai, and C. Qiu, "A correlation analysis

- method for power systems based on random matrix theory,” *ArXiv e-prints*, Jun. 2015, accepted by IEEE Trans on Smart Grid. [Online]. Available: <http://arxiv.org/pdf/1506.04854.pdf>
- [20] —, “A correlation analysis method for operation status of distribution network based on random matrix theory,” *Proceedings of the CSEE*, vol. 40, no. 3, pp. 781–790, Mar. 2016.
 - [21] Y. Yan, G. Sheng, H. Wang, Y. Liu, Y. Chen, X. Jiang, and Z. Guo, “The key state assessment method of power transmission equipment using big data analyzing model based on large dimensional random matrix,” *Proceedings of the CSEE*, vol. 36, no. 2, pp. 435–445, Jan. 2016.
 - [22] F. Götze and A. Tikhomirov, “The rate of convergence for spectra of gue and lue matrix ensembles,” *Open Mathematics*, vol. 3, no. 4, pp. 666–704, 2005.
 - [23] A. J. Hey, S. Tansley, K. M. Tolle *et al.*, *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research Redmond, WA, 2009, vol. 1.
 - [24] J. Gray, “Jim gray on escience: A transformed scientific method,” *The fourth paradigm: Data-intensive scientific discovery*, pp. xvii–xxxi, 2009.
 - [25] R. Kitchin, “Big data and human geography opportunities, challenges and risks,” *Dialogues in human geography*, vol. 3, no. 3, pp. 262–267, 2013.
 - [26] M. Shcherbina, “Central limit theorem for linear eigenvalue statistics of the wigner and sample covariance random matrices,” *ArXiv e-prints*, Jan. 2011. [Online]. Available: <http://arxiv.org/pdf/1101.3249.pdf>
 - [27] C. Zhang and R. C. Qiu, “Massive mimo as a big data system: Random matrix models and testbed,” *IEEE Access*, vol. 3, pp. 837–851, 2015.
 - [28] R. Qiu and P. Antonik, *Smart Grid and Big Data*. John Wiley and Sons, 2015.
 - [29] J. R. Ipsen and M. Kieburg, “Weak commutation relations and eigenvalue statistics for products of rectangular random matrices,” *Physical Review E*, vol. 89, no. 3, 2014, Art. ID 032106.
 - [30] F. Benaych-Georges and J. Rochet, “Outliers in the single ring theorem,” *Probability Theory and Related Fields*, pp. 1–51, May 2015. [Online]. Available: <http://dx.doi.org/10.1007/s00440-015-0632-x>
 - [31] T. Tao, “Outliers in the spectrum of iid matrices with bounded rank perturbations,” *Probability Theory and Related Fields*, vol. 155, no. 1-2, pp. 231–263, 2013.